Prediksi, Diagnosis dan Pengobatan Acute Myloid Leukemia Menggunakan Big Data Analisis

H. Hamrul*, A. Irianti
Universitas Sulawesi Barat
Email: *heliawatyhamrul@unsulbar.ac.id

Abstrak

Penyakit kanker adalah penyakit utama yang telah menjadi ancaman terbesar bagi kesehatan manusia karena deteksi dini sulit dilakukan, diagnosis dan biaya perawatan yang relatif mahal. Salah satu jenis kanker yang banyak diderita Acute Myeloid Leukemia. Kanker ini adalah salah satu jenis kanker darah yang mengakibatkan sumsum tulang belakang tidak dapat menghasilkan sekelompok sel darah putih seri myeloid yang matang. Orang yang menderita Acute Myeloid Leukemia sulit dideteksi secara dini sebab gejalanya yang mirip dengan gejala flu. Menurut survei Organisasi Kesehatan Dunia Tahun 2012 8,2 juta kasus kematian terkait penyakit ini. Oleh sebab itu, penelitian mengenai kanker darah menjadi topik utama dalam bidang medis dan bioinformatika dan terus berkembang hingga saat ini, termasuk teknologi big data dan penerapan algoritma untuk memprediksi, mendiagnosis dan mengobati penyakit tersebut. Dalam bidang kesehatan, big data digunakan untuk memprediksi penyakit, menganalisis gejala, meningkatkan akurasi diagnosis, menyediakan obat yang tepat bagi pasien, meningkatkan kualitas perawatan, menurunkan biaya pengobatan dan meningkatkan rentang hidup dan mengurangi dampak kematian. Kemajuan teknologi big data dapat dimanfaatkan untuk menyelamatkan pasien dan mengurangi resiko kematian pasien kanker melalui deteksi dini. Literature review ini bertujuan untuk mengkaji hasil penelitian mengenai bagaimana big data dapat digunakan memprediksi, mendiagnosis dan mengobati acute myeloid leukemia. Metode penelitian dilakukan dengan cara mengekslorasi sumber data dari tiga database utama, Scopus, ScienceDirect, DOAJ, EBSCO, Web of Science yang mengindeks jurnal dan prosiding conference yang diterbitkan oleh IEEE, ACM, SpringerLink, dan Elsevier. Artikel yang dipilih adalah artikel yang terbit 5 tahun terakhir (2014-2019). Hasil yang diperoleh yakni terdapat beberapa algoritma statistik yang digunakan untuk memprediksi, mediagnosis dan mengobati Acute myeloid leukemia diantaranya algoritma MERGER, knearest neighbor (k-NN), decision tree (DT), Support Vector Machine (SVM) dan Surveillance, Epidemiology, and End Results (SEER), dan Algoritma Map Reduce. Algoritma ini kemudian diimpementasikan pada Hadoop Framework. Big data analysis dapat digunakan untuk memprediksi, mendiagnosis dan memberikan informasi pemberian obat yang tepat bagi penderita Acute myeloid leukemia.

Kata kunci: *big data analisis, algoritma, acute myeloid leukemia*

1. Pendahuluan

Penyakit kanker adalah penyakit utama yang telah menjadi ancaman terbesar bagi kesehatan manusia karena deteksi dini sulit dilakukan, diagnosis dan biaya

perawatan yang relatif mahal. Lebih dari 100 jenis kanker membutuhkan cara diagnosis dan pengobatan khusus. Salah satu jenis kanker darah yang banyak diderita Acute Myeloid Leukemia. Kanker ini adalah salah satu jenis kanker darah yang mengakibatkan sumsum tulang belakang tidak dapat menghasilkan sekelompok sel darah putih seri myeloid yang matang. Orang yang menderita Acute Myeloid Leukemia sulit dideteksi secara dini sebab gejalanya yang mirip dengan gejala flu. Menurut survei Organisasi Kesehatan Dunia Tahun 2012 8,2 juta kasus kematian terkait penyakit ini. Oleh sebab itu, penelitian mengenai kanker darah menjadi topik utama dalam bidang medis dan bioinformatika dan terus berkembang hingga saat ini, termasuk teknologi big data dan penerapan algoritma untuk memprediksi, mendiagnosis dan mengobati penyakit tersebut. Dalam bidang kesehatan, big data digunakan untuk memprediksi penyakit, menganalisis gejala, meningkatkan akurasi diagnosis, menyediakan obat yang tepat bagi pasien, meningkatkan kualitas perawatan, menurunkan biaya pengobatan dan meningkatkan rentang hidup dan mengurangi dampak kematian. Kemajuan teknologi big data dapat dimanfaatkan untuk menyelamatkan pasien dan mengurangi resiko kematian pasien kanker melalui deteksi dini.

Karakteristik Big Data

Volume: berkaitan dengan ukuran data yang super besar.

Variety: menggambarkan tipe atau jenis data yang sangat beragam yang meliputi berbagai jenis data baik data yang terstruktur maupun data yang tidak terstruktur dalam suatu dabase maupun data yang tidak terorganisir dalam suatu dalbase seperti halnya data teks data *webpages*, data suara, video, *click stream, log file* dan lain sebagainya.

Velocit: dapat diartikan sebagai kecepatan dihasilkannya suatu data dan seberapa cepat data itu harus diproses agar dapat memberikan hasil yang valid.

Perangkat Analisis Big Data

Apache Hadoop: adalah suatu *framework* yang memungkinkan pemrosesan secara terdistribusi terhadap data yang berukuran besar dengan melibatkan sat atau lebih kluster komputer dan menerapkan *programming model* yang sederhana. *Framework* Hadoop terdiri atas 4 komponen utama yakni:

a. *Hadoop Distribused File System* (HDFS) adalah file sistem terdistribusi yang memfasilitasi penyimpanan data secara terdistribusi dalam kluster komputer.
b. *Hadoop Mapreduce*, adalah sebuah sistem yang ditujukan untuk memproses data berukuran besar secara paralel.
c. *Hadoop Common* adalah common utilitas yang digunakan untuk mendukung modul modul Hadoop yang lainnya.
d. Hadoop YARB adalah *framework* yang berperan dalam *job scheduling* dan *resource management* pada kluster Hadoop. Hadoop versi 1 tidka memliki YARN. Dilihat dari fungsinya, YARN juga dianggap sebagai sistem operasi Hadoop.

Spark SQL: Spark SQL adalah seperangkat alat dari API yang mendukung *DataFrames* yang mirip dengan Python tetapi yang berjalan di atas dataset yang terdistribusi walaupun tidak memiliki semua fungsi yang serupa Spark SQL mirip Pyton.

2. Metode

Metode penelitian dilakukan dengan cara mengekslorasi sumber data dari tiga database utama, Scopus, ScienceDirect, DOAJ, EBSCO, Web of Science yang mengindeks jurnal dan prosiding conference yang diterbitkan oleh IEEE, ACM, SpringerLink, dan Elsevier. Artikel yang dipilih adalah artikel yang terbit 5 tahun terakhir (2014-2019**).** Paper yang dipilih merupakan paper yang membahas mengenai big data, algoritma machine learning, dan leukemia khususnya Acute Myeloid Leukemia. Makalah ini merupakan literature review yang mencoba mengkaji penerapan *big data* dan algoritma *machine learning* yang dapat digunakan untuk memprediksi, mendiagnosis dan memberikan pengobatan yang tepat pada penderita *acute myeloid leukemia*

3. Hasil dan Pembahasan

a. Prediksi Acute Myeloid Leukemia

Mendioagnisis Acute Myeloid Leukemia sangat bergantung pada data klinis. Prediksi Acute Myeloid Leukemia bisa membantu para ahli medis melalui data klinis pasien. Industri perawatan kesehatan misalnya rumah sakit, mengumpulkan data-data atau lebih dikenal sebagai big data dari industry kesehatan lainnya. Setelah data terkumpul maka dilakukan penambangan data atau *data mining* untuk menemukan pola-pola data yang berkaitan dengan penyakit pasien sehingga dapat memberikan hasil prediksi dan diagnosis yang akurat. Saat ini, sebagian besar rumah sakit menyortir data dan informasi yang dimiliki untuk memanajemen data pasien melalui sistem manajemen data pasien. Sistem ini menghasilkan data dalam jumlah besar yang meliputi teks, gambar, bagan dan angka. Data data tersebut dapat digunakan untuk memprediksi, mendiagnosis dan membantu pengobatan Acute Myeloid Leukemia dengan cara data dianalisis menggunakan beberapa algoritma misalnya Naive Bayes, Neural Network dan Decision Tree. Hasil analisis ini memberikan hasil yang akurat dan dapat digunakan untuk membantu pengambilan keputusan bagi para medis.

b. Algoritma Machine Learning

1) Naive Bayes Classifier

Algoritma pertama untuk memprediksi Acute Myeloid Leukemia adalah algoritma Naive Bayes. dalam algoritma ini, 13 atribut *preprocessed* digunakan sebagai data inputan. Semua atribut independen atau tidak terpengaruh oleh atribut lainnya, dan secara signifikan mengurangi perhitungan. Rumus Naive Bayes sebagai berikut:

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)} \quad (1)$$

$$P\ (\ c \mid X\ ) =\ P\ (x_1 \mid c)\ x\ P\ ((x_2 \mid c)\ x \ldots x\ P\ ((x_n \mid c)\ x\ P(\ c\ ) \quad (2)$$

Dimana:

Fp(c|x) - probabilitas posterior kelas (target) yang diberikan prediktor (atribut)

P(c) - prior probabilitas kelas, juga disebut prior. Ini adalah peluang mengamati kelas secara umum.

P (x|c) - kemungkinan yang merupakan probabilitas prediktor yang diberikan kelas
P (x) - probabilitas prediktor sebelumnya juga disebut pembuktian

Maka dengan input catatan pasien dari 13 atribut kita dapat menghitung probabilitas posterior untuk semua tingkat risiko yang mungkin. Pasien memiliki tingkat risiko yang posterior probabilitas nya maksimum. Training data set digunakan untuk perhitungan probabilitas bersyarat kelas. Diberikan atribut xi kita dapat menghitung P (xi | Cj) untuk kelas Cj. Untuk ini, kita dapat menggunakan definisi dasar probabilitas.

$$P(x_i|Cj) = \frac{\text{Number of times } x_i \text{ occurs in rows of training data set } X^t \text{ for clas}}{\text{Number of times Cj occurs in training data set } X^t} \quad (3)$$

xi € X dimana j=0,1,2,3,4. Untuk perhitungan kemungkinan, eluruh dataset training digunakan. Namun metode perhitungan ini hanya berlaku jika variabel-variabel tersebut bersifat diskrit seperti, tubuh lemas, nyeri sendi dan tulang. untuk data rekam medis pasien. Dalam dataset terdapat 5 atribut utama digunakan yaitu, usia, kejang, mimisan, pendarahan, penglihatan, keseimbangan dan bersifat bersifat kontinu. Oleh karena itu digunakan fungsi kepadatan probabilitas untuk pendekatan awal, perhitungan kepadatan kondisional kelas menggunakan asumsi distribusi normal untuk semua variabel kontinu seperti yang ditunjukkan berikut ini:

$$p(x = v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}} \quad (4)$$

Menggunakan metode distribusi normal untuk gejala pendarahan dan pembengkakan lebih cocok digunakan. akan tetapi untuk gejala memar dan penglihatan kabur tidak cocok menggunakan metode ini. sebagai hasilnya, didapatkan hasil yang sebagian akurat dan sebagian tidak, Kita dapat menggunakan pendekatan lain untuk menghindari distribusi variabel, yaitu dengan asumsi variabel menjadi diskrit. Dalam kasus seperti itu, perhitungan probabilitas bersyarat kelas untuk variabel-variabel ini dilakukan dengan cara yang sama seperti yang dilakukan untuk variabel diskrit lainnya. Nilai asumsi ini berlaku untuk kasus ini karena mengandung kumpulan data yang besar dan juga mengarah pada hasil yang diharapkan dan akurasi yang tinggi.

2) Probabilistic Analiysis and Clasification (PAC)

Algoritma *machine learning* yang berasal dari Algoritma Naive Bayes adalah *Probabilistic Analysis and Classification* (PAC). Algoritma Ini menggunakan konsep perhitungan probabilitas rata-rata dari seluruh data set training {Xt}. Algoritma Ini dibentuk dari model Naive Bayes untuk mengatasi kelemahan algoritma Naive Bayes. Salah satu keuntungan adalah menggunakan teknik diskritisasi. Dalam teknik ini dilakukan pengurangan lengkap variabel kontinu menjadi variabel diskrit. Keuntungan lain adalah karena konversi lengkap variabel kontinu ke variabel diskrit. Metode *Laplacian Smooting* yang digunakan dalam klasifikasi Naive Bayes tidak diperlukan, yang pada gilirannya mengurangi perbandingan yang tidak perlu dan instruksi yang tidak diinginkan.

Konsep utama dari algoritma ini adalah menggunakan perhitungan rata-rata tertimbang untuk semua nilai penyakit leukemia sampai dan kecuali ditemukan tuple

yang sama persis dalam data set. Jadi dalam hal ini tingkat risiko tuple ditetapkan ke tingkat risiko input di rekam medis pasien. Kasus ini sangat jarang terjadi sehingga harus menggunakan perhitungan rata-rata tertimbang untuk seluruh kumpulan data dan menghitung kontribusi masing-masing dan setiap nilai untuk tingkat risiko tertentu dan menemukan solusi untuk kontribusi berbeda untuk seluruh kumpulan data. Dalam mempertimbangkan seluruh data set, digunakan sejumlah tupel pendukung untuk berbagai tingkat risiko dalam kumpulan data. Konsep ini mirip dengan *"Prior"* dalam algoritma Naive Bayes tetapi dalam algoritma Naive Bayes probabilitas sebelumnya memberikan bobot lebih ke tingkat risiko berdasarkan nilai-nilai variabel. Dalam PAC, konsep ini secara bersamaan mengurangi berat ini yang menghasilkan kesalahan, karena perbedaan peningkatan persentase pembilang dan penyebut dalam αi. Jadi untuk mengatasi kerugian ini, kami mengalikannya dengan faktor normalisasi untuk mengurangi kesalahan dan memberikan hasil yang akurat. Akhirnya istilah maksimum μi di antara semua tingkat risiko dikembalikan sebagai tingkat risiko pasien

c. Hadoop Map Reduce Programming untuk Pemrosesan Data

Berdasarkan analisis algoritma yang telah dilakukan maka diperolah hasil bahwa algoritma PAC merupakan algoritma yang sangat baik dalam memberikan prediksi risiko penyakit acute myeloid leukemia dengan tingkat akurasi yang tinggi. Algoritma PAC ini dibangun menggunakan algoritma *machine learning* yang menutupi kelemahan dari algoritma sebelumnya dan meningkatkan akurasi prediksi tingkat risiko penyakit. Banyak rumah sakit dan industri perawatan kesehatan memiliki data pasien dalam jumlah besar. Dengan pertumbuhan populasi yang luar biasa, para dokter dan para ahli yang tersedia kurang proporsional dengan populasi di mana para dokter kadang-kadang gagal untuk mendiagnosis dengan benar tingkat keparahan penyakit. Hadoop, sebuah cluster node digunakan untuk memproses Big Data. Proses Map Reduce diterapkan untuk algoritma yang dirancang.

Mapper: Di dalam fungsi *Mapper* setiap baris dari file input diambil sebagai input ke fase map dan dibawa ke *map-task* yang berbeda secara paralel, dengan mempertimbangkan *multi-node cluster* karena setiap node mengikuti prosedur yang sama secara bersamaan. Jika ada N baris dalam file input dan memiliki nilai *default* M *map-tasks* maka jumlah baris yang diproses oleh masing-masing *map task*, maka fungsi *mapper* mengeksekusi algoritma PAC dan setiap *map task* node dan setiap node dalam *multi node cluster*. Setiap saat *mapper* membutuhkan satu baris dari *big data* sebagai input dan proses algoritma *machine learning* untuk menghitung tingkat risiko. Tapi, di sini nomor baris diambil sebagai kunci dan seluruh baris diambil sebagai nilai. Tingkat risiko disediakan sebagai kunci dan nilai diberikan ke setiap atribut yang perlu dievaluasi. File konteks adalah output antara yang diberikan oleh fungsi *mapper* sebagai input ke fungsi pengurang.

Reducer: Reducer mengacak tingkat risiko yang disediakan oleh file konteks dan mengurutkannya sesuai dengan nilai-nilai kunci yang diberikan pada fungsi *reducer* berdasarkan urutan dan menyimpan output yang diurutkan dalam file. *Map reducer* digunakan untuk memproses Big Data di kedua jenis algoritma. Berbagai fungsi *map reducer* diterapkan untuk menghitung grafik untuk atribut yang berbeda dengan jumlah pasien penderita leukemia dengan dan tanpa penyakit.
Map dan *Reduce* adalah sebuah fungsi yang keduanya didefinisikan sehubungan dengan data terstruktur dalam pasangan (kunci, nilai). Map mendefinisikan satu pasangan data

dengan sebuah jenis domain data, dan mengembalikan daftar pasangan di domain data yang berbeda

$$Map(k1,v1) \rightarrow list(k2,v2)$$

Fungsi *Map* diterapkan secara paralel ke setiap pasangan (kunci dengan k1) dalam dataset yang berisi input. Fungsi ini menghasilkan daftar pasangan (dikunci oleh k2) untuk setiap panggilan. Setelah itu, metode *MapReduce* selesai mengumpulkan semua pasangan dengan kunci yang sama (k2) dari semua daftar data dan mengelompokkannya bersama sebagai *cluster*, membuat satu grup untuk setiap kunci. Fungsi *Reduce* kemudian dilakukan secara paralel ke setiap grup, yang menghasilkan kumpulan nilai dalam domain yang sama:

$$Reduce(k2,list(v2)) \rightarrow list(v3)$$

Reduce menghasilkan satu nilai atau nol, meskipun satu panggilan bisa mengembalikan lebih dari satu nilai.

d. Hasil Analisis Grafis

Output dari analisis big data ini dapat berupa laporan pasien untuk input berbasis form atau output grafis jika file Big Data disediakan sebagai input. Sebuah studi perbandingan algoritma machine learning yang dijelaskan sebelumnya dibuat dan grafik akurasi diplot untuk menentukan algoritma terbaik untuk prediksi penyakit. Ini termasuk beberapa aspek penelitian seperti jumlah total pasien yang memiliki dan tidak memiliki penyakit Leukemia, jumlah pasien dari usia tertentu yang memiliki dan tidak memiliki penyakit dll. Semua aspek ini ditampilkan dalam format grafis sehingga lebih mudah dipahai oleh pengguna. Tabel 1 berikut ini menunjukkan studi perbandingan algoritma Machine Learning seperti yang dijelaskan dalam makalah ini Naive Bayes untuk variabel kontinu (89.90), Naive Bayes untuk variabel diskrit (95.21) dan Algoritma PAC (97,48).

Tabel 1. Perbandingan Hasil Analisis Algoritma

Machine Learning Algorithms	Accuracy
NAÏVEBAYES CONTINUOUS VARIABLE	89.80%
NAÏVE BAYES DISCRETE VARIABLE	95.21%
PROBABILISTIC ANALYSIS	97.48%

e. Membangun Sistem Terpusat dan Tersebar pada Cloud Platform HDPS

Cloud computing adalah proses yang melibatkan distribusi jaringan komputer, di mana suatu program atau aplikasi dapat berjalan pada banyak komputer yang terhubung pada saat yang sama. *Cloud Computing* Ini secara khusus mengacu pada sebuah mesin perangkat keras komputer atau kumpulan mesin perangkat keras komputer yang biasanya disebut sebagai server yang terhubung melalui jaringan komunikasi seperti Internet, juga jaringan seperti (LAN) atau (WAN). Setiap pengguna individu yang memiliki izin untuk mengakses server dapat menggunakan sumber daya server untuk menjalankan aplikasi, dan juga untuk menyimpan data atau melakukan tugas komputasi lainnya. Proyek ini dikembangkan pada *Platform Cloud* yang disebut Jelastic.

Jelastic adalah *platform* yang melibatkan karakteristik *as Infrastructure* layanan *cloud computing* yang menyediakan solusi jaringan, server, dan penyimpanan untuk

klien pengembangan perangkat lunak, bisnis perusahaan, OEM, dan penyedia hosting web. eberapa perusahaan telah mengembangkan teknologi menggunakan Java dan PHP ke *platform* berbasis *cloud*.

Jelastik memiliki mitra hosting internasional dan pusat data. Perusahaan dapat menyediakan fasilitas seperti memori, CPU, dan ruang disk untuk memenuhi kebutuhan pelanggan. Pesaing utama Jelastic adalah Google App Engine, Amazon Elastic Beanstalk, Heroku, dan Cloud Foundry. Jelastic adalah *platform* unik yang tidak memiliki batasan atau persyaratan perubahan kode, dan juga menawarkan penskalaan vertikal otomatis, manajemen siklus aplikasi, dan ketersediaan dari berbagai penyedia host di seluruh dunia.

4. Kesimpulan

Data terkait perawatan kesehatan sangat banyak dan berasal dari berbagai sumber yang yang berbeda. Data data ini sebagian besar berupa data tidak terstruktur. Saat ini, prediki, diagnosis dan pengobatan *Acute Myleoid Leukemia* dilakukan dengan cara memanfaatkan pengetahuan dan pengalaman para ahli medis dan screening data medis dari pasien terkumpul didatabase selama proses diagosis.

Penerapan algoritma *machie learning* yang akurat digunakan untuk memprediksi, mendiagnosis dan membantu memberikan pengobatan yang pada penyakit *acute myeloid leukemia*. dan perbandingan algoritma dilakukan untuk mengevaluasi akurasi menggunakan grafik. Lebih mudah untuk memahami grafik dan pengguna juga dapat menentukan tingkat risikonya dan untuk mendapatkan laporan yang serupa. Algoritma ini dapat digunakan pada ayanan *cloud* dan *big data* dapat dengan mudah diakses dan diproses.

Referensi

[1] M. Kumar, K. R. Nitish and K. R. Santanu “Analysis of Microarray Leukemia Data Using an Efficient MapReduce Based K-Nearest Neighbor Classifier,” *Journal of Biomedical Informatics,* , vol. 60, no. 7, pp. 395-409, 2016.

[2] A. S. Khaled, F. M. Wael, Y. A. A. Maghari "Proceeding ICIT " in *International Conference on Information Technology*, 2017.

[3] S. I. Lee and S. Celik, “A Machine Learning Approach To Integrate Big Data For Precision Medicine In Acute Myeloid Leukemia”, *Journal of Natture Communication*, pp. 9-42, 2018.

[4] C. Bruno, Medeiros, S. S. Hoang, D. Hurst and K. Q. Hoang, “Big data analysis of treatment patterns and outcomes among elderly acute myeloid leukemia patients in the United States,” *Journal of Big Data,* , no 94, pp. 1127-1138, 2015.

[5] P. A. Rahayuningsih, “Komparasi Algoritma Klasifikasi Data Mining untuk Memprediksi Tingkat Kematian Dini Kanker dengan Dataset Early Death Cancer ”*Journal of Information Technology and Computer Science,* , vol. 4, no. 2, pp. 65-68, 2019.

[6] D. R. Umesh and B. Ramachandra, “Big Data Analytics to Predict Breast Cancer Recurrence on SEER Dataset using MapReduce Approach,” *International Journal of Computer Application*, vol. 150, no. 7, pp. 7-11, 2016.